Acquisition d'une grammaire catégorielle depuis un corpus d'arbre en français

Eric Poupard[1], Denis Béchet[2], Annie Foret[3]

[1] IRISA - Université de Rennes 1 - France
eric_poupard@hotmail.com

[2] LINA - Université de Nantes - France
Denis.Bechet@univ-nantes.fr

[3] IRISA - Université de Rennes 1 - France
Annie.Foret@irisa.fr

Les grammaires catégorielles ont montré leur intérêt dans le domaine du traitement du langage naturel. Dans l'optique de l'acquisition automatique de grammaires, leur caractère lexicalisé est un avantage important comme l'ont souligné de nombreux articles sur le sujet en particulier l'article présentant l'algorithme RG (Buszkowski & Penn, 1990) sur lequel est basé ce travail.

Cet algorithme est difficilement utilisable directement. D'une part, les grammaires résultats sont supposées être rigides, c'est-à-dire ne posséder qu'un seul type pour tous les mots du lexique ce qui n'est pas le cas avec les langues naturelles. D'autre part, les phrases en entrée de l'algorithme doivent être présentées sous la forme de *structures FA*, c'est-à-dire d'arbres binaires dont les nœuds internes sont étiquetés soit par l'*application en avant*, soit par l'*application en arrière* et les feuilles par les mots de la phrase. Ceci nécessite un corpus de structures FA qui n'existe pas pour le français.

Ce problème a été abordé pour l'anglais dans le cadre des grammaires catégorielles combinatoires (CCG) (Hockenmaier & Steedman, 2002). Dans cet article, les auteurs transforment les arbres de la Penn Treebank en structures FA puis utilisent un algorithme d'apprentissage ad-hoc pour en déduire un lexique de l'anglais.

Notre solution reprend les deux phases présentés dans cet article : transformation du corpus d'arbres en *structures FA annotées* puis application d'un algorithme apprentissage sur ces structures pour obtenir le lexique. Elle s'en distingue sur deux points importants. Premièrement, nous traitons le français et pas l'anglais en utilisant un corpus d'arbres syntaxiques de phrases françaises issues du journal "Le Monde" d'une taille d'un million de mots et développé à l'Université Paris 7 (Abeillé *et al.*, 2003). Ceci impose une approche différente de la phase de transformation des arbres en structures FA. Deuxièmement, nous avons basé notre algorithme d'apprentissage sur l'algorithme RG. La différence principale par rapport à l'algorithme RG porte sur l'abandon de la phase terminale d'unification des catégories qui permet d'obtenir une seule catégorie pour chaque mot mais qui n'est pas utilisable dans le contexte du français. En fait, nous avons remplacer cette phase d'unification par une utilisation des catégories présentes

draft

sur les arbres du corpus initial ce qui justifie l'utilisation de structures FA annotées plutôt que de structures FA simples.

La phase de transformation en structures FA annotées qui est une des originalités de ce travail détermine pour chaque constituant syntaxique composé (par exemple un groupe nominal) quel est l'élément principal et quels sont ses arguments et ses modificateurs. Voici un exemple de structure FA annotée obtenue :

S :$\backslash_e$[**GN** :$/_e$[`L'`, **NC** :`opération`],
$/_e$[`prendra`,
GN :[`deux`, **NC** :`formes`]]

H : pour *Head*
C : pour *Complement*
M : pour *Modifier*
S : pour *Sentence*
GN : pour *Groupe Nominal*
NC : pour *Nom Commun*

L'algorithme d'apprentissage est basé sur l'algorithme RG. Nous avons supprimé la phase finale d'unification des types d'un même mot. Par contre, nous avons tenu compte du fait que les structures d'entrées ne sont pas des structures FA mais des structures FA annotées par des sous-constituants. Voici un extrait du lexique engendré :

Mot	Type
expliquer	$\_121401/GN$
exploitants	NC
explosions	NC
explosives	$NC\backslash GNI$
exportations	NC
exprimé	$\_420591\backslash\_420525$
exprimé	$\_69136\backslash((GN\backslash S)/GN)$
expulsion	NC
extraordinaire	$NC\backslash GNI$

Le résultat final est assez intéressant car le lexique engendré contient relativement peu de variables de type engendré par l'algorithme et beaucoup de types primitifs correspondant à des catégories fixées à priori (NC pour les noms communs, GN pour les groupes nominaux, GNI pour les groupes nominaux sans article, etc).

TODO : conclusion

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–188. Kluwer, Dordrecht.

BUSZKOWSKI W. & PENN G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*, **49**, 431–454.

HOCKENMAIER J. & STEEDMAN M. (2002). Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*.

POUPARD E. (2005). Apprentissage de grammaires catégorielles à partir d'exemples annotés foncteur-argument. Master's thesis, Mémoire de DEA informatique, Université Rennes I.

Draft